ISSN 1561-5359. Штучний інтелект, 2016, № 4
УДК 004.93

М.М. Сажок, Р.А. Селюх, Д.Я. Федорин, О.А. Юхименко
Міжнародний науково-навчальний центр інформаційних технологій та систем, Україна
пр. Академіка Глушкова, 40, м. Київ, 03680

ТЕКСТОНЕЗАЛЕЖНЕ РОЗПІЗНАВАННЯ ДИКТОРІВ ІЗ ЗАСТОСУВАННЯМ ІНШОМОВНОГО КОРПУСУ

M.M. Sazhok, R.A. Seliukh, D.Y. Fedoryn, O.A. Yukhymenko
International Research and Training Centre for Information Technologies and Systems, Ukraine
40, Akademician Hlushkov av., 40, Kyiv, 03680

TEXT-INDEPENDENT SPEAKER RECOGNITION USING A FOREIGN LANGUAGE SPEECH CORPORA

У статті описано експериментальне дослідження з використання іншомовного корпусу для текстонезалежного розпізнавання дикторів. Цей підхід дав би змогу, за відсутності мовленнєвих ресурсів для оцінки параметрів, застосувати великий мовленнєвий корпус з іншої мови. Як іншомовний корпус використано відкритий для вільного доступу мовленнєвий корпус THUYG-20 SRE. Наведено результати досліджень.

Ключові слова: текстонезалежне розпізнавання дикторів, *i*-вектори, ймовірнісний лінійний дискримінаційний аналіз.

This paper presents an experimental research on text-independent speaker recognition using a foreign language speech corpus. Such a technique would allow for involving a large speech corpus when no sufficient speech data available for a given language. Autors considered the open and free speech database THUYG-20 SRE as a foreign corpus. Research results are discussed.

Key words: text-independent speaker recognition, i-vectors, probabilistic linear discriminant analysis (plda).

Вступ

Метою розпізнавання диктора є необхідність ідентифікувати, якому диктору належить мовленнєвий сигнал. З 90-х років домінував підхід GMM-UBM (Gaussian Mixture Model - Universal Background Model), але зараз провідним підходом є використання *i*-векторів (*i-vectors*). Цей підхід передбачає двоступеневе оцінювання параметрів. По-перше - оцінюються параметри загальної UBM, у рамках GMM-UBM, параметри моделі *i*-вектору та параметри порівняння двох векторів, наприклад, на основі PLDA. По-друге - оцінюються параметри індивідуальних моделей дикторів.

Загальновідомі експериментальні дослідження показали, що для створення загальної моделі необхідно мати великий мовленнєвий корпус, де представлено якомога більше різних дикторів, що говорять заданою мовою. Для створення *i*-векторів конкретних дикторів достатньо й декількох десятків секунд їхнього мовлення. Автори вирішили перевірити, наскільки зміниться надійність розпізнавання, коли для створення загальної моделі береться мовленнєвий корпус не тією мовою, якою спілкуються диктори, яких планується розпізнавати, а деякою іншою мовою. Як іншомовний корпус автори обрали *THUYG*-20 *SRE* - уйгурський мовленнєвий корпус, у ролі дикторів для тестування - україномовний корпус *UkReco*. Вибір уйгурського мовленнєвого корпусу зумовлений тим, що він єдиний відомий корпус, який знаходиться у вільному доступі, містить інформацію про стать диктора та для якого доступні скрипти, що дають змогу оцінювати параметри для розпізнавання дикторів на основі інструментальних засобів *Kaldi* [1].

Метою роботи є експериментальне дослідження можливості використання

іншомовного корпусу для розпізнавання дикторів за мовленнєвим сигналом українською мовою.

Задача розпізнавання дикторів та шляхи її розв'язання

В останні роки найкращі результати у розпізнаванні дикторів показав підхід *i*-векторів [2].

Для заданої фрази модель *i*-вектора припускає, що дикторозалежний супервектор генерується таким чином:

$$M = m + T w \quad (1),$$

де m - супервектор, незалежний від диктора та каналу мовлення, T - матриця з низьким рангом та w - низьковимірний вектор, що представляє вимовлену фразу. Припускаючи, що w має нормальний розподіл $N(0, I)$, рівняння (1) розглядається як лінійна гауссівська модель, оцінюваний параметр M має гауссівський розподіл $N(m, T\,T^{T})$. Оцінка параметрів та виведення змінних може бути виконано стандартним чином. На підставі мовленнєвого сигналу $\{X_i\}$ навчальної вибірки, матриця T оцінюється оптимізацією такої функції ймовірності:

$$L(T)=\sum_i \ln\{P(X_i;T)\}=\sum_i \ln\left\{\sum_M P(X_i;M)P(M;T)\right\} \quad (2),$$

де умовна імовірність $P(X_i;M)$ моделюється сумішшю нормальних законів, а апріорна ймовірність $P(M;T)$ - гауссоїд. Коли матрицю T оцінено, обчислення постеріорної ймовірності для w у фразі X не викликає ускладнень, оскільки $P(w|X)$ є також гауссоїдом. Лише вектор середніх значень (так званий *i*-вектор) обчислюється за допомогою оцінки апостеріорного максимуму (*MAP - maximum aposteriori probability*).

За мовленнєвим сигналом, представленим *i*-векторами, ймовірність спостереження деякого прогнозованого диктора за умов тестового сигналу обчислюється як косинус-відстань між *i*-векторами тестового сигналу та навчального сигналу прогнозованого диктора.

Модель *i*-вектору — це модель тотальної варіабельності (*total-variability*). Це означає, що *i*-вектори представляють характеристики як диктора, так і акустичного каналу. Ймовірнісний лінійний дискримінантний аналіз (PLDA) розділяє *total-variability* простір на підпростір диктора та підпростір каналу. Тому диктори можуть бути представлені більш точно. Ця модель може бути сформульована як:

$$w_r = m + Ux_r + Vy + e_r \quad (3)$$

де w_r - це *i*-вектор r-ї фрази, m - середнє значення сукупності, U - підпростір каналу, x_r - вектор каналу, V - підпростір диктора, y - вектор диктора, e_r - похибка. Параметри x_r та y мають стандартний гауссівський розподіл, а e_r - гауссівський розподіл $N(0, \Sigma)$. Параметри $\{m, U, V, \Sigma\}$ оцінюються з використанням алгоритму максимізації математичного сподівання (*EM - expectation maximization*), а виведення для вектору диктора y, як правило, досягається засобами MAP.

Розпізнавання дикторів уйгурською мовою

Експерименти з розпізнаванням дикторів уйгурською мовою були проведені в університеті Цінхуа [3]. Для розпізнавання був використаний корпус *THUYG*-20 *SRE*. Записи для нього були виконані в офісі на вуглецевий мікрофон. Частота дискретизації — 16 *KHz*. Усі диктори - студенти віку 19-28 років, походять із 30

районів Китаю. Дикторам давали читати різноманітну літературу загальної тематики.

Таблиця 1. Характеристика корпусу уйгурською мовою

	Диктори	Жінки	Чоловіки	Фрази	Обсяг (годин)
Загальна модель	200	100	100	4771	13.15
Індивідуальні моделі	153	87	66	153	1.28
Тестування дикторів	153	87	66	2361	6.56

Базова система розпізнавання дикторів була побудована на технології *i*-векторів, яка включає модель і-вектору для диктора та ряду способів порівняння векторів. Для опису мовленнєвого сигналу використовувалися 20-розмірні мел-частотні кепстральні коефіцієнти (*MFCC*) та їх перша та друга похідні (усього розмірність - 60). Для усунення ефекту спотворень, що вносяться акустичним каналом, застосовувалася кепстральна нормалізація середнього та дисперсії (*CMVN*). UBM містила 2048 гауссоїдів та *i*-вектор розмірністю 400. Навчання на диктора проводилося на 10, 20 та 30 секундах.

Наведена нижче таблиця містить результати запуску скрипту, який міститься з базою уйгурського мовлення.

Таблиця 2. Результати розпізнавання дикторів: *EER* для корпусу THUYG-20

Метод порівняння векторів	Жінки			Чоловіки		
	10с	20с	30с	10с	20с	30с
cosine	8.4	6.3	4.8	10.7	9.1	7.6
lda	7.1	5.3	3.7	6.3	5.6	4.9
plda	5.3	3.9	3.1	6.2	5.3	4.4

Слід зауважити, що корпус *THUYG*-20 має обсяг відносно невеликий для задачі розпізнавання дикторів - 13.15 годин. Тому було додатково проведено експерименти з використанням англомовного корпусу *Fisher* (219.59 годин, жіноча частина) [4], як доповнення до корпусу для оцінки параметрів загальної моделі.

Таблиця 3. Результати розпізнавання дикторів: *EER* для корпусу THUYG-20 з додаванням корпусу *Fisher*

Тренувальна БД	Навчальна БД	10с	20с	30с
THUYG-20 *SRE*	*THUYG*-20 *SRE*	6.35	5.11	4.01
THUYG-20 *SRE + Fisher*	*THUYG*-20 *SRE*	5.03	2.92	2.33

Цей експеримент, з одного боку, демонструє те, що задача розпізнавання дикторів не має виключної прив'язаності до мови, а з іншого - підтверджує припущення про необхідність значних обсягів мовленнєвого сигналу для побудови загальної моделі.

Розпізнавання дикторів українською мовою

У дослідженнях ми використали україномовний багатодикторний мовленнєвий корпус *UkReco*, який містить понад 30 000 реалізацій слів і тисячі речень близько 100 дикторів, що мешкають у різних областях України. Реалізації слів зберігають частотні пропорції фонем і є фонетично збалансованими, при підборі слів також

враховувалися їх частотні характеристики [5]. Цей мовленнєвий корпус було створено завдяки гранту Президента України для обдарованої молоді, контракт № 32 від 30.05.2006 р.

З мовленнєвого корпусу взято до розгляду матеріал, записаний з голосу 63 дикторів (40 жінок і 23 чоловіки). Цю основну вибірку було розділено на дві частини. Частина файлів кожного диктора була взята для навчання (10, 20 та 30 секунд), а частина - для розпізнавання.

Слід зауважити, що на відміну від уйгурського корпусу, в якому при тестуванні використовувалися злиті фрази тривалістю 10 секунд, із *UkReco* були взяті сегменти, що містять окремо вимовлені слова, тривалість у середньому кожного сегменту - 1.5 секунд.

Таблиця 4. Характеристики вибірок, що використані в роботі з українською мовою

	Диктори	Жінки	Чоловіки	Фрази	Обсяг (годин)	Мова
Загальна модель	200	100	100	4771	13.15	Українська
Індивідуальні моделі	63	40	23	63	0.54	Українська
Тестування	63	40	23	630	0.35	Українська

Таблиця 5. Результати розпізнавання дикторів: *EER* для корпусу *UkReco*

Метод порівняння векторів	Жінки			Чоловіки		
	10с	20с	30с	10с	20с	30с
cosine	32	28	27.8	33.9	32.2	31.3
lda	27	23.5	24	20.9	17.4	16.5
plda	19	17.5	18.5	15.2	14.8	14.8

Параметри, за яких отримано найменшу похибку, відповідають відомим результатам [3]. Втім, у нашому випадку, спостерігається більша похибка при розпізнаванні дикторів-жінок.

Висновки

Дане дослідження показало, що в разі недостатньої кількості мовленнєвих ресурсів при оцінюванні параметрів універсальної моделі для систем розпізнавання дикторів доцільно використовувати іншомовний мовленнєвий корпус. Зазначимо, що уйгурська мова належить до тюркських мов, а українська - до слов'янських. Отже, слушним є припущення, що при мовах більш близьких результати були би ще кращими.

Література

1. Povey D. The Kaldi speech recognition toolkit / Povey D. et al. // ASRU, 2011.
2. Dehak N. Front-End Factor Analysis for Speaker Verification / N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet // IEEE Trans. Audio, Speech & Language Processing 19(4): 788-798, 2011.
3. Rozi A. An open/free database and benchmark for uyghur speaker recognition / A. Rozi, D. Wang, Z. Zhang // Oriental COCOSDA, 2015.
4. Cieri C. The Fisher corpus: a resource for the next generations of speech-to-text / C. Cieri, D. Miller, K. Walker // 4th IC on LRE, 2004.
5. Сажок М. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови / М. Сажок, Р. Селюх, О. Юхименко // Штучний інтелект. - 2009. - №4. - С. 230-233.

Literatura

1. Povey D. The Kaldi speech recognition toolkit / Povey D. et all // ASRU, 2011.
2. Dehak N. Front-End Factor Analysis for Speaker Verification / N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet // IEEE Trans. Audio, Speech & Language Processing 19(4): 788-798, 2011.
3. Rozi A. An open/free database and benchmark for uygur speaker recognition / A. Rozi, D. Wang, Z. Zhang // Oriental COCOSDA, 2015.
4. Cieri C. The Fisher corpus: a resource for the next generations of speech-to-text / C. Cieri, D. Miller, K. Walker // 4th IC on LRE, 2004.
5. Sazhok M. Adaptatsiya akustychnykh modeley fonem do holosu dyktora dlya pofonemnoho rozpiznavannya izol'ovanykh sliv ukrayins'koyi movy / M. Sazhok, R. Selyukh, O. Yukhymenko // Shtuchnyy intelekt. - 2009. - #4. - S. 230-233.

RESUME

M.M. Sazhok, R.A. Seliukh, D.Y. Fedoryn, O.A. Yukhymenko

Text-Independent Speaker Recognition Using Another Language Speech Corpora

This paper presents an experimental research on text-independent speaker recognition based on briefly described i-vectors technique. Conventionally, to build a state-of-the-art experimental system a large speech corpus is necessary to estimate parameters for the universal model. Since such a Ukrainian corpus has not been accessible for the authors, using a foreign language speech corpus was considered. Authors selected the open and free speech database THUYG-20 SRE as a foreign language corpus. Presented experimental research shows promising results. The best results corresponds to probabilistic linear discriminant analysis technique while comparing i-vectors extracted from the input signal with a speaker model. The applied technique confirmed prospectiveness for involving a foreign language large speech corpus when no sufficient speech data available for the desired language.

Надійшла до редакції 18.10.2016